W416–W419 *Nucleic Acids Research, 2007, Vol. 35, Web Server issue*
doi:10.1093/nar/gkm228

3dLOGO: a web server for the identification, analysis and use of conserved protein substructures

Allegra Via[1,*], Daniele Peluso[1], Pier Federico Gherardini[1], Emanuele de Rinaldis[1,2], Teresa Colombo[3,4], Gabriele Ausiello[1] and Manuela Helmer-Citterich[1]

[1]Centre for Molecular Bioinformatics, Department of Biology, University of Rome Tor Vergata, 00133 Rome, Italy, [2]Bioinformatics Group, I.R.B.M. P. Angeletti, MRL-Rome, Via Pontina Km, 30600 Pomezia, Italy, [3]Center for Comparative Functional Genomics, Department of Biology, New York University, NY 10003, USA and [4]Systems Biology Group - Max-Delbrück-Centrum für Molekulare Medizin, Berlin

Received December 21, 2006; Revised March 15, 2007; Accepted March 28, 2007

ABSTRACT

3dLOGO is a web server for the identification and analysis of conserved protein 3D substructures. Given a set of residues in a PDB (Protein Data Bank) chain, the server detects the matching substructure(s) in a set of user-provided protein structures, generates a multiple structure alignment centered on the input substructures and highlights other residues whose structural conservation becomes evident after the defined superposition. Conserved residues are proposed to the user for highlighting functional areas, deriving refined structural motifs or building sequence patterns. Residue structural conservation can be visualized through an expressly designed Java application, 3dProLogo, which is a 3D implementation of a sequence logo. The 3dLOGO server, with related documentation, is available at http://3dlogo.uniroma2.it/

INTRODUCTION

The superimposition of a set of related protein structures is a straightforward methodology for identifying structurally conserved residues that cannot be detected using sequence alignments (1,2). Structurally conserved regions are usually of functional significance and can also improve our understanding of protein evolution. Despite the fact that a variety of multiple structure alignment algorithms have been reported over the years [(3–9); for a comprehensive review see (10)], methods for automatic identification of variable and conserved regions within a multiple structure alignment of proteins have not been fully exploited. The approach of identifying structurally conserved residues in a set of protein structures has been already used (11) to improve the sensitivity and/or specificity of poorly performing PROSITE patterns (12). Clearly, the construction of structure and sequence motifs is an important area, which would greatly benefit from the possibility of easily detecting residues structurally conserved in diverse protein structures. The 3dLOGO server makes accessible to web users a generalized and automated tool for the detection of 3D conserved residues in a multiple structure alignment (MStA), for the interactive improvement of an MStA, for obtaining a pictorial view of a 3D consensus and for deriving a sequence motif from a structural consensus.

Starting from a set of protein structures and a number of specified residues on at least one of them (i.e. a substructure), 3dLOGO identifies the substructures common to the set of the input structures and performs a 3D alignment onto the substructures' residues. A tabular view is presented highlighting, if present, other similar and well-superimposed amino acids. The output set of conserved residues can be used for several purposes: to recognize a putative functional region, to build a 3D pattern for structure database searches, to identify event(s) of convergent evolution, to derive a customized sequence pattern in the form of a regular expression, which is usable for sequence database searches. Interesting results can be obtained with a careful selection of the input residues: such as a set of amino acids known to specifically bind the same ligand in different protein structures [e.g. some of the p-loop residues in proteins that bind ATP or GTP (2)] or also residues belonging to a PROSITE pattern true positive match in diverse protein structures (11).

METHODS

Overview

The 3dLOGO web server combines three procedures. First, a multiple alignment of protein structures is built on the substructures common to a set of

*To whom correspondence should be addressed. Tel: +39 067259 4324; Fax: +39 067259 4314; Email: allegra.via@uniroma2.it
Correspondence may also be addressed to Manuela Helmer-Citterich. Email: citterich@uniroma2.it

user-selected structures. Then the alignment is analyzed to identify conserved residues; the structural alignment can be further improved by iterating the whole procedure and using the newly identified conserved residues as input. Finally, a sequence consensus can be built with the spatially conserved residues that are co-linear in the corresponding protein sequences.

A complete description of the method and usage technical details can be found on the documentation web pages.

The input data

Users must supply n ($n \geq 2$) protein chains, which can be specified using PDB codes (13) or whose coordinates can be uploaded. For reasons of speed, the maximum number of input structures is 25. For each given structure, a single chain ID must be provided. The user must also supply the number (ID) of at least three amino acids of at least one input protein chain. This minimal substructure (the 'object' substructure) cannot comprise more than ten residues and is needed in order to localize the protein area to be studied. For instance, the minimal substructure can be a region known to bind a ligand (e.g. the solvent exposed residues of an SH3 domain or a set of residues binding ATP or GTP), to have a specific function (e.g. the catalytic triad of a serine protease), to match a known functional motif (e.g. PROSITE) or it can be any other protein area of interest. The server uses the Query3d application (14) for the detection of local similarities, in order to identify a structural match of the 'object' substructure (query) in the other $n-1$ input structures. A match is valid if (a) it is made up of at least three residues, (b) pairs of matching residues are physico-chemically similar and display an averaged rmsd <2 Å. If no local similarities are found in one of the $n-1$ input structures, this last is skipped and the subsequent steps are carried out only for the remaining $n-2$ structures. The 'residue similarity options' in the input page, makes it possible to set the degree of physico-chemical similarity under which 3D conserved positions are identified, e.g. by setting 'identical', only positions made up of identical residues will be shown.

Multiple alignment of structures

All structures are superimposed using two points for each given residue, namely the C-alpha and a pseudo-atom, calculated as the average of the coordinates of the residue side-chain atoms. In the case of n input structures, the first structure is assumed to be the *reference* (or 'target') structure for the structural superimposition(s). The remaining $n-1$ structures are aligned pairwise to the reference structure.

The output data: identification of conserved residues

After superimposing the input structures on the given residues, all the nearby conserved residues are identified. Two amino acids belonging to different structures are assumed to be conserved if they are spatially well superimposed and display physico-chemical similarities, according to a substitution matrix [see (15) and the documentation web page for more details]. The user can make a selection from the input page if only identical, very similar, similar or all the spatially well-superimposed residues have to be displayed. Identification of the conserved residues is carried out using the 3D profiles method (1). The multi-alignment output (Figure 1) is displayed in a table containing a column for each input structure and a row for each conserved position. User-supplied residues are highlighted by green dots, whereas yellow marks pinpoint newly identified well-superimposed residues.

In the score column, the similarity measure derived from the substitution matrix is shown. Its value is maximum if the row's residue types are identical. The 'rmsd' column reports the average rmsd between the residues belonging to a row of the table with respect to the reference structure. Only positions whose residues display an average rmsd <2.5 Å are shown in the table. The structural alignment can be visualized with the Astex Viewer™ applet (16) or downloaded in text format (Figure 1). Furthermore, the user can visualize the variable or conserved positions of the structural multiple alignment through a specifically designed Java application, **3dProLogo**, which is a 3D implementation of a sequence logo (17).

The 3dProLogo Java application

A sequence logo is a graphical representation of an amino acid or nucleic acid multiple sequence alignment (MSA) which provides an instant way of visualizing variable and conserved positions of the MSA (17). In the sequence logo, the logo is constructed by calculating the information content of each position of the aligned sequences, and then displaying the characters representing the amino acids stacked on top of each other. The height of each letter is made proportional to its frequency. The height of each stack is then adjusted according to the information content at that position, as detailed in (17). Similarly, 3dProLogo is a 3D graphical representation of the residues in a multiple structural alignment. The total information for each position in the structural alignment is calculated as in (17) and used to derive the height of each residue type's letter present in the 3D position. The resulting representation (3D logo) can be seen in Figure 1, where the conserved positions in P-loop containing proteins 3D multiple alignment (see the last section) are displayed. The strong conservation of the residues involved in the nucleotide binding can be detected immediately as can be their arrangement in space. 3D logos can be rotated according to the users need.

Sequence consensus building

In order to build a sequence consensus, the user is allowed to manually select the positions in the MStA that better fit with his/her requirements. When the user is satisfied with the proposed structural alignment, a sequence *consensus* (pattern) can be generated from the matched residues. By using all the identified conserved residues, the sequence pattern is constructed as a regular expression for searching sequence databases. The format of the regular expression

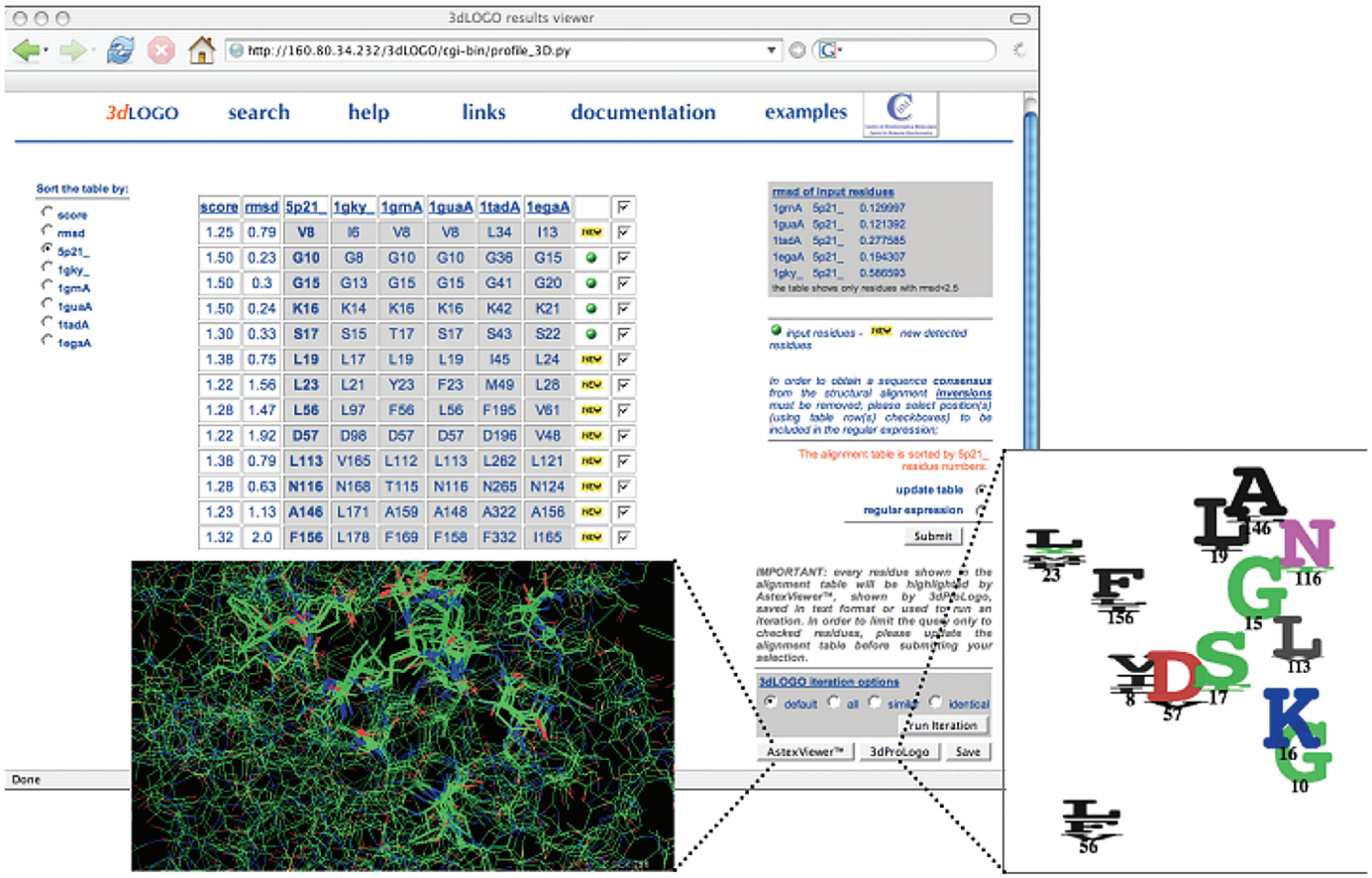

Figure 1. The 3dLOGO output page. The AstexViewer window appears when clicking on the corresponding button at the bottom of the web page. The residues of the alignment table are highlighted in the viewer. The conserved positions in the 3D multiple alignment can be displayed with the 3dProLogo viewer by clicking on the 3dProLogo button. The output page shown in the figure was obtained using the input default parameter for residue–residue similarity, which means that only the positions with similarity score >1.2 are displayed.

provided is the one required for direct submission to the ScanProsite (18) server.

Iterated alignments

The search for structurally conserved residues can be repeated by superimposing the structures on a new choice of conserved residues, including the newly detected ones. The whole process can be iterated until the identification of an 'extended' structure or sequence 'consensus'. The user can obtain a new alignment by starting from a manual selection of the conserved positions listed in the structural alignment output table and using these as a new input of the procedure.

ONE EXAMPLE: P-LOOP CONTAINING PROTEINS

The P-loop is a phosphate-binding loop commonly found in ATP and GTP-binding proteins (2,19). The 'consensus' sequence of this loop is [AG]XXXXGK[TS] (20). A typical example of a P-loop-containing GTP-binding protein is the small GTPase c-H-ras p21, belonging to the Ras family (SwissProt: P01112). Guanylate kinase (SwissProt: Q8UGD7) is an ATP-binding protein belonging to the guanylate kinases family. The two proteins display 11% sequence identity, and the SSM (21) structure alignment of the two proteins finds only a few highly similar matching residues (concentrated around the P-loop), and an overall lack of structural similarity. The P-loop in the c-H-ras p21 structure (PDB 5P21) matches residues 10–18, whereas in the guanylate kinase structure (PDB: 1gky) the P-loop is found at residues 8-16 (2). We considered 5p21 and 1gky plus a set of four different P-loop-containing proteins (PDB IDs: 1grn,1gua,1tad,1ega) extracted from (2) to perform a 3dLOGO query. By using these six structures as 3dLOGO input, all the residues belonging to the P-loop were found, as well as several new well-superimposed residues which are close to the P-loop in space, but very distant along the sequence (Figure 1). In particular we retrieved the ras N116 (correctly superimposed to the kinase N168), which belongs to the well-known G protein conserved motif NKXD, whose residues are involved in the interaction with the nucleotide.

More examples are described in detail and graphically displayed in the 3dLOGO examples Web page (http://3dlogo.uniroma2.it/3dLOGO/examples.html).

CONCLUSIONS

The 3dLOGO web server is designed for the local study of protein structures. It represents a new useful tool for the analysis of conserved structural regions and for the derivation of more specific structure and sequence patterns. The 3dLOGO tool was also implemented to produce a novel 3D representation of the conservation of residues in a set of protein structures. Representative examples are reported in the Results and in the 3dLOGO Web pages.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of Telethon (GGP04273), a PNR 2001-2003 (FIRB art.8) and a PNR 2003-2007 (FIRB art.8). We thank Memmo Federici for his essential contribution to the 3dProLogo applet. Funding to pay the Open access publication charges for this article was provided by AIRC.

Conflict of interest statement. None decalerd

REFERENCES

1. de Rinaldis,M., Ausiello,G., Cesareni,G. and Helmer-Citterich,M. (1998) Three-dimensional profiles: a new tool to identify protein surface similarities. *J. Mol. Biol.*, **284**, 1211–1221.
2. Via,A., Ferre,F., Brannetti,B., Valencia,A. and Helmer-Citterich,M. (2000) Three-dimensional view of the surface motif associated with the P-loop structure: cis and trans cases of convergent evolution. *J. Mol. Biol.*, **303**, 455–465.
3. Leibowitz,N., Nussinov,R. and Wolfson,H.J. (2001) MUSTA–a general, efficient, automated method for multiple structure alignment and detection of common motifs: application to proteins. *J. Comput. Biol.*, **8**, 93–121.
4. Dror,O., Benyamini,H., Nussinov,R. and Wolfson,H. (2003) MASS: multiple structural alignment by secondary structures. *Bioinformatics*, **19**(Suppl. 1), 95–104.
5. Guda,C., Lu,S., Scheff,E.D., Bourne,P.E. and Shindyalov,I.N. (2004) CE-MC: a multiple protein structure alignment server. *Nucleic Acids Res.*, **32**, W100–W103.
6. Maiti,R., Van Domselaar,G.H., Zhang,H. and Wishart,D.S. (2004) SuperPose: a simple server for sophisticated structural superposition. *Nucleic Acids Res.*, **32**, W590–W594.
7. Ochagavia,M.E. and Wodak,S. (2004) Progressive combinatorial algorithm for multiple structural alignments: application to distantly related proteins. *Proteins*, **55**, 436–454.
8. Shatsky,M., Nussinov,R. and Wolfson,H.J. (2004) A method for simultaneous alignment of multiple protein structures. *Proteins*, **56**, 143–156.
9. Lupyan,D., Leo-Macias,A. and Ortiz,A.R. (2005) A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics*, **21**, 3255–3263.
10. Lemmen,C. and Lengauer,T. (2000) Computational methods for the structural alignment of molecules. *J. Comput. Aided Mol. Des.*, **14**, 215–232.
11. Via,A. and Helmer-Citterich,M. (2004) A structural study for the optimisation of functional motifs encoded in protein sequences. *BMC Bioinformatics*, **5**, 50.
12. Hulo,N., Sigrist,C.J.A., Le Saux,V., Langendijk-Genevaux,P.S., Bordoli,L., Gattiker,A., De Castro,E., Bucher,P. and Bairochi,A. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res.*, **32**, 134–137.
13. Deshpande,N., Addess,K.J., Bluhm,W.F., Merino-Ott,J.C., Townsend-Merino,W., Zhang,Q., Knezevich,C., Xie,L., Chen,L. *et al.* (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.*, **33**, D233–D237.
14. Ausiello,G., Via,A. and Helmer-Citterich,M. (2005) Query3d: a new method for high-throughput analysis of functional residues in protein structures. *BMC Bioinformatics*, **6** (Suppl. 4), S5.
15. Schwartz,R.M. and Dayhoff,M.O. (1979).Matrices for detecting distant relationships. In: Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, Washington, DC, pp. 353–359.
16. Hartshorn,M.J. (2002) AstexViewer: a visualisation aid for structure-based drug design. *J. Comput. Aided Mol. Des.*, **16**, 871–881.
17. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acid Res.*, **18**, 6097–6100.
18. Gattiker,A., Gasteiger,E. and Bairoch,A. (2002) ScanProsite: a reference implementation of a PROSITE scanning tool. *Applied Bioinformatics.*, **1**, 107–108.
19. Vetter,I. and Wittinghofer,A. (1999) Nucleoside triphosphate-binding proteins: different scaffolds to achieve phosphoryl transfer. *Quart. Rev. Biophys.*, **32**, 1–56.
20. Saraste,M., Sibbald,P.R. and Wittinghofer,A. (1990) The P-loop – a common motif in ATP- and GTP-binding proteins. *Trends Biochem. Sci.*, **15**, 430–434.
21. Krissinel,E. and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr.*, **60** (Pt 12 Pt 1), 2256–2268.